LASSO transductif et autres généralisations

Pierre Alquier, Mohamed Hebiri

LASSO transductif et autres généralisations

Pierre Alquier[(1)] & Mohamed Hebiri[(2)]

(1,2) LPMA, Université Paris 7, 175 rue du Chevaleret, 75013 Paris, FRANCE
(1) CREST-ENSAE, 3, avenue Pierre Larousse, 92240 Malakoff, FRANCE

Abstract: We consider the linear regression problem, where the number p of covariates is possibly larger than the number n of observations, under sparsity assumptions. In this work, we propose a generalized version of the LASSO by (Tibshirani, 1996), based on geometrical remarks about the LASSO provided by Alquier and Hebiri (2008) and that takes into acount the objective of the statistician. As a special case, we consider the problem of estimating the regression vector in the transductive setting as described by Vapnik (1998), in which the estimator construction is based on a new unlabelled dataset of interest. From a theoretical point of view, we derive Sparsity Inequalities (SI) for our estimator, i.e., bounds on the estimation error involving the sparsity of the parameter we try to estimate. We also derive a *Pathwise Coordinate Optimization* algorithm to provide an approximated solution for our estimator.

Résumé : On considère le problème de régression linéaire dans lequel le nombre de variables explicatives p peut être plus grand que le nombre d'observation n. Sous des hypothèses de parcimonie, nous proposons dans cette étude une généralisation de l'estimateur LASSO de Tibshirani (1996), qui s'appuie sur des considérations géométriques présentées par Alquier and Hebiri (2008) et prenant en compte l'objectif du statisticien. Le problème de l'estimation du paramètre inconnu dans le cadre transductif (Vapnik, 1998) est également considéré, i.e., une approche dans laquelle la construction de l'estimateur s'appuie sur un nouvel échantillon non étiqueté et pour lequel nous souhaitons réaliser de bonne performances de prédiction. Du point de vue théorique, nous illustrons nos résultats par des "Inégalités de Sparsité", i.e., des bornes sur l'erreur d'estimation qui font intervenir la parcimonie du paramètre que l'on veut estimer. Nous proposons également un algorithme d'optimisation coordonnée par coordonnée pour approximer notre estimateur.
Mots Clés : Modèles de régression; Choix de modèles; Parcimonie; LASSO.

1 Introduction

Dans un certain nombre d'applications (imagerie, puces ADN,...), le statisticien peut avoir à sa disposition des données très volumineuses. Certains problèmes de régression, en particulier, présentent un nombre de variables p important, parfois supérieur à la taille n de l'échantillon. Dans cette situation, un objectif important est la sélection d'un petit nombre de variables pertinentes. Dans ce but, un certain nombre de méthodes de régression ont été proposées, depuis les critères d'informations classiques comme AIC

par Akaike (1973) et BIC par Schwartz (1978), jusqu'à des méthodes de régularisation plus récentes comme celles basées sur la norme ℓ_1 : le LASSO par Tibshirani (1996), ou le *Dantzig selector* par Candès et Tao (2007). Ces dernières méthodes ont reçu une attention particulière lors de ces dernières années, en partie dû à la possibilité de les implémenter rapidement même pour de grandes valeurs de p.

Plus formellement, introduisons le modèle de régression suivant :

$$y_i = x_i\beta^* + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

où le design $x_i = (x_{i,1}, \dots, x_{i,p}) \in \mathrm{R}^p$ est déterministe, $\beta^* = (\beta_1^*, \dots, \beta_p^*)' \in \mathrm{R}^p$ est un paramètre inconnu et les $\varepsilon_1, \dots, \varepsilon_n$ sont i.i.d., gaussiens centrés et de variance σ^2. L'objectif est alors de faire de l'inférence sur des quantités faisant intervenir le paramètre β^*. Soit X la matrice dont la i-ème ligne est x_i, et notons X_j sa j-ème colonne, avec $i \in \{1, \dots, n\}$ et $j \in \{1, \dots, p\}$. On a donc $X = (x_1', \dots, x_n')' = (X_1, \dots, X_p)$. Pour simplifier les notations, on supposera que les obesrvations sont renormalisées de façon à avoir $X_j'X_j/n = 1$. On pose aussi $Y = (y_1, ..., y_n)'$. Pour tout $d \in \mathrm{N}$, tout $v \in \mathrm{R}^d$ et tout $\alpha \geq 1$ on utilisera la notation habituelle $\|v\|_\alpha = (|v_1|^\alpha + ... + |v_d|^\alpha)^{1/\alpha}$, en particulier $\|\cdot\|_2$ sera la norme euclidienne.

Entre autres, le LASSO (noté ici $\hat{\beta}_L$) a été proposé comme estimateur valable pour traiter ce problème quand p est grand, et même plus grand que n, avec de très bons résultats pratiques : des simulations et des tests sur des données réelles peuvent être trouvés, par exemple, dans Tibshirani (1996). D'un point de vue théorique, des "Inégalités de Sparsité" (le terme français n'est pas consacré, on dira donc simplement SI, pour *Sparsity Inequality*, dans la suite) ont été prouvées pour ces estimateurs : adoptons la notation $|\cdot|$ pour désigner le cardinal d'un ensemble; ainsi des bornes sur $(1/n)\|X\hat{\beta}_L - X\beta^*\|_2^2$ ou sur $\|\hat{\beta}_L - \beta^*\|_2^2$, de la forme $cste.\sigma^2|\{j \in \{1, ..., p\}, \beta_j^* \neq 0\}| \log(p)/n$, faisant intervenir le nombre de coordonnées non nulles dans β^* (multiplié par $\log(p)$), au lieu de p, sont énoncées et prouvées, par exemple par Bickel, Ritov et Tsybakov (2007). Des conditions assurant que $\hat{\beta}_{LASSO}$ et β^* ont le même ensemble de coordonnées non nulles sont proposées par Bunea (2008).

Remarquons que les travaux mentionnés donnent des bornes sur $\|X\hat{\beta}_L - X\beta^*\|_2^2$ ou $\|\hat{\beta}_L - \beta^*\|_2^2$. Ils assurent donc que, sous certaines hypothèses, $X\hat{\beta}_L$ est un bon estimateur de $X\beta^*$, ou que $\hat{\beta}_L$ est un bon estimateur de β^*. Cependant, supposons que l'on donne au statisticien des observations supplémentaires, $x_i \in \mathrm{R}^p$ pour $n+1 \leq i \leq n+m$ (notons Z la matrice $(x_{n+1}', \dots, x_{n+m}')'$) et que l'objectif du statisticien soit précisément d'estimer $Z\beta^*$ (dans l'absolu, il ne cherche pas une bonne estimation de β^*, mais simplement à deviner les labels Y_i associés aux nouveaux x_i). Dans la monogarphie de Vapnik (1998), une argumentation en faveur d'une approche dédiée à la résolution de ce problème est développée : un estimateur, dit transductif pour estimer $Z\beta^*$ doit être proposé. Celui-ci peut être différent de l'estimateur utilisé pour estimer β^* ou $X\beta^*$. Ainsi des méthodes de classification supervisée ont été étendues avec succès au cas transductif; on mentionnera par exemple les *Support Vector Machines* (SVM) par Vapnik (1998) et l'estimateur de

Gibbs par Catoni (2008). Le livre de Chapelle, Schölkopf et Zien (2006) sur la classification semi-supervisée apporte également des explications sur l'effet de stabilisation sur les estimateurs, que peut produire la prise en compte de l'information apportée par les nouveaux x_i.

Dans la Partie 2, on propose un estimateur général $\hat{\beta}_{M,\lambda}$ (2) qui dépend de deux paramètres de réglage: λ, un paramètre de régularisation, et une matrice M de taille $p \times p$, qui va permettre de calibrer l'estimateur en fonction de l'objectif du statisticien. En particulier, suivant le choix de la matrice M (discuté plus en détail plus loin), on pourra considérer les trois objectifs suivants : premièrement, le **débruitage ou estimation de** $X\beta^*$, l'estimateur considéré est dans ce cas le LASSO; deuxièmement la **transduction, ou estimation de** $Z\beta^*$ et nous appelerons l'estimateur associé $\hat{\beta}_{M,\lambda}$: LASSO transductif; finalement l'**estimation directe de** β^* et pour cette perspective, l'estimateur proposé ne sera correctement défini que si $p < n$ et est alors une version seuillée de l'estimateur des moindres carrés ordinaires (MCO). L'estimateur (2) est en fait tirée d'une version plus complète de ce travail, Alquier et Hebiri (2008), qui s'appuie sur les considérations géométriques sur le LASSO discutées par Alquier (2008). Dans la Partie 3, on donne une "Inégalité de Sparsité" pour $\hat{\beta}_{M,\lambda}$ (Théorème 3.1).

Enfin, des algorithmes pour calculer $\hat{\beta}_{LASSO}$ ou $\hat{\beta}_{DS}$ sont connus. Ainsi, pour le LASSO, nous pouvons citer les méthodes de point intérieur de Kim, Koh, Lustig, Boyd et Gorinevsky (2007) par exemple, l'algorithme LARS de Efron, Hastie, Johnstone et Tibshirani (2004) ou encore les algorithmes d'optimisation coordonnée par coordonnée (ou *Pathwise Coordinate Optimization*, PCO dans la suite) discutés par Friedman, Hastie, Höfling et Tibshirani (2007). On montre dans la Partie 4 qu'un algorithme PCO peut facilement être proposé pour approcher $\hat{\beta}_{M,\lambda}$, notons toutefois que les arguments théoriques en faveur des algorithmes PCO sont en général insuffisants, et que l'on a pour le moment pas de garanties théoriques sur la convergence de cet algorithme.

2 Définition de l'estimateur $\hat{\beta}_{M,\lambda}$

Pour $\lambda > 0$ et toute matrice symétrique positive M, on considère l'estimateur

$$\hat{\beta}_{M,\lambda} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ -2Y'X(\widetilde{X'X})^{-1}M\beta + \beta' M\beta + 2\lambda \|\Xi_M \beta\|_1 \right\}, \tag{2}$$

où $(\widetilde{X'X})^{-1} = (X'X)^{-1}$ si $(X'X)$ est inversible, et n'importe quelle pseudo-inverse de cette matrice sinon, et où Ξ_M est une matrice diagonale $p \times p$ dont le j-ème coefficient est $\xi_j^{\frac{1}{2}}(M)$ avec $\xi_j(M) = \frac{1}{n}[M(\widetilde{X'X})^{-1}M]_{j,j}$. Par la suite, on va considérer trois cas particuliers de cet estimateur de manière adaptée à l'objectif :

- **débruitage :** on prend $\hat{\beta}_{X'X,\lambda}$, qui est donné par

$$\begin{aligned}\hat{\beta}_{X'X,\lambda} &\in \arg\min_{\beta\in\mathrm{R}^p}\left\{\|Y-X\beta\|_2^2+2\lambda\|\beta\|_1\right\}\\ &= \arg\min_{\beta\in\mathrm{R}^p}\left\{-2Y'X\beta+\beta'X'X\beta+2\lambda\|\beta\|_1\right\},\end{aligned}$$

  on reconnaît donc l'estimateur LASSO (remarquons qu'ici $\Xi_{X'X}=I$);

- **transduction :** on appellera "LASSO transductif" l'estimateur $\hat{\beta}_{\frac{n}{m}Z'Z,\lambda}$ donné par

$$\hat{\beta}_{\frac{n}{m}Z'Z,\lambda}\in\arg\min_{\beta\in\mathrm{R}^p}\Big\{-\frac{2n}{m}Y'X(\widetilde{X'X})^{-1}(Z'Z)\beta+\frac{n}{m}\beta'Z'Z\beta+2\lambda\|\Xi_{\frac{n}{m}Z'Z}\beta\|_1\Big\};$$

- **estimation de β^* :** on prendra $\hat{\beta}_{nI,\lambda}$, defini par

$$\hat{\beta}_{nI,\lambda}\in\arg\min_{\beta\in\mathrm{R}^p}\Big\{-2Y'X(\widetilde{X'X})^{-1}\beta+\beta'\beta+2\lambda\|\Xi_I\beta\|_1\Big\}.$$

Proposition 2.1 *Supposons que $(X'X)$ soit inversible. Alors $\hat{\beta}_{nI,\lambda}$ est l'estimateur des MCO seuillé : à savoir, si $\hat{\beta}_{MCO}=(X'X)^{-1}X'Y$ alors $\hat{\beta}_{nI,\lambda}$ est obtenu en remplaçant la j-ème coordonnée $b_j=(\hat{\beta}_{MCO})_j$ de $\hat{\beta}_{MCO}$ par $\mathrm{Sgn}(b_j)\left(|b_j|-\lambda\xi_j(nI)/n\right)_+$, où $\forall u\in\mathrm{R}$, $\mathrm{Sgn}(u)=\pm1$ selon le signe de u et 0 si $a=0$ et où $(u)_+=\max\{u,0\}$.*

3 Résultats théoriques

Définition 3.1 *On dit que la matrice M de taille $p\times p$ satisfait l'Hypothèse $A(M)$ s'il existe $c(M)>0$ tel que pour tout $\alpha\in\mathrm{R}^p$ vérifiant $\sum_{j:\beta_j^*=0}\xi_j(M)\,|\alpha_j|\le 3\sum_{j:\beta_j^*\neq0}\xi_j(M)\,|\alpha_j|$ on ait*

$$\alpha'M\alpha\ge[c(M)/n]\,\alpha'\alpha.$$

Commentons rapidement cette hypothèse. Dans le cas facile, où M est inversible, la condition $\alpha'M\alpha\ge[c(M)/n]\,\alpha'\alpha$ est satisfaite pour tout $\alpha\in\mathrm{R}^p$ avec $c(M)/n$ égal à la plus petite valeur propre de M. Cependant, prenons pas exemple le cas du LASSO, $M=(X'X)$ qui ne peut être inversible si $p>n$. Dans ce cas, l'hypothèse peut toutefois être vérifiée, car elle n'exige pas que $\alpha'M\alpha\ge[c(M)/n]\,\alpha'\alpha$ soit vrai pour tout $\alpha\in\mathrm{R}^p$ mais seulement pour un petit sous-ensemble de R^p. Pour $M=(X'X)$, cette hypothèse est exactement celle prise par Bickel, Ritov et Tsybakov (2007).

Théorème 3.1 *Supposons que l'Hypothèse $A(M)$ soit satisfaite. Supposons que $Ker(M)=Ker(X)$. Choisissons $0<\varepsilon<1$ et $\lambda=2\sigma\sqrt{2n\log\left(\frac{p}{\varepsilon}\right)}$. Avec probabilité au moins $1-\varepsilon$ sur le tirage de Y, on aura*

$$\left(\beta^*-\hat{\beta}_{M,\lambda}\right)'M\left(\beta^*-\hat{\beta}_{M,\lambda}\right)\le\frac{72\sigma^2}{c(M)}\sum_{j:\beta_j^*\neq0}\xi_j^2(M)\log\left(\frac{p}{\varepsilon}\right).$$

La preuve est donnée par Alquier et Hebiri (2008). En particulier, on obtient pour les différents objectifs :

- si $A(X'X)$ est satisfaite, avec probabilité au moins $1-\varepsilon$,

$$\frac{1}{n}\left\|X\left(\hat{\beta}_{X'X,\lambda}-\beta^*\right)\right\|_2^2 \leq \frac{72\sigma^2}{nc(X'X)}\left|\{j:\beta_j^*\neq 0\}\right|\log\left(\frac{p}{\varepsilon}\right);$$

- si $A(\frac{n}{m}Z'Z)$, et si $Ker(Z)=Ker(X)$, avec probabilité au moins $1-\varepsilon$,

$$\frac{1}{m}\left\|Z\left(\hat{\beta}_{\frac{n}{m}Z'Z,\lambda}-\beta^*\right)\right\|_2^2 \leq \frac{72\sigma^2}{nc(\frac{n}{m}Z'Z)}\sum_{j:\beta_j^*\neq 0}\xi_j\left(\frac{n}{m}Z'Z\right)\log\left(\frac{p}{\varepsilon}\right);$$

- si $(X'X)$ est inversible, avec probabilité au moins $1-\varepsilon$,

$$\left\|\hat{\beta}_{nI,\lambda}-\beta^*\right\|_2^2 \leq \frac{72\sigma^2}{n}\sum_{j:\beta_j^*\neq 0}\xi_j(nI)\log\left(\frac{p}{\varepsilon}\right).$$

Remarque 3.1 *Donc, sous certaines hypothèses, chaque estimateur satisfait une SI pour l'objectif pour lequel il a été conçu.*

Remarque 3.2 *Pour $M=(X'X)$, les résultats obtenus sont similaires à ceux de Bunea, Tsybakov et Wegkamp (2007). Par ailleurs, la preuve du Théorème 3.1 donnée par Alquier et Hebiri (2008) repose sur une généralisation d'arguments donnés par ces auteurs.*

4 Algorithme PCO

Remarquons que la Proposition 2.1 donne un moyen simple de calculer l'estimateur $\hat{\beta}_{nI,\lambda}$ en pratique. Quand au LASSO, il existe un grand nombre de méthodes pour le calculer en pratique (méthodes de points intérieurs, LARS, algorithmes PCO,...). La question du calcul de $\hat{\beta}_{\frac{n}{m}Z'Z,\lambda}$ se pose donc. Dans la sous-partie suivante, on propose un algorithme PCO pour le calcul de $\hat{\beta}_{M,\lambda}$ en général ($M=(X'X)$ redonne l'algorithme PCO étudié par Friedman, Hastie, Höfling et Tibshirani (2007)). Insistons sur le fait qu'on ne prouve pas ici que cet algorithme converge vers $\hat{\beta}_{M,\lambda}$ de façon générale.

Algorithme : L'algorithme part de $\beta \leftarrow (0,...,0)'$. Puis, jusqu'à ce que β se stabilise, on parcourt les coordonnées ($j=1,...,p$) en prenant

$$\beta_j = \mathrm{Sgn}(a_j)\left(|a_j|-\frac{\lambda\xi_j^{\frac{1}{2}}(M)}{M_{j,j}}\right)_+,$$

où a_j est défini par $a_j=\frac{\left(Y'X(\widetilde{X'X})^{-1}M\right)_j-\sum_{\ell\neq j}\beta_\ell M_{i,j}}{M_{j,j}}$.

5 Conclusion

Inspiré par un travail plus détailé de Alquier et Hebiri (2008), on a proposé un estimateur général $\hat{\beta}_{M,\lambda}$ qui a pour cas particulier le LASSO et une nouvelle version du LASSO permettant de résoudre le problème de la transduction. Un prolongement de ce travail consisterait à montrer que l'algorithme proposé dans la Partie 4 converge effectivement vers notre estimateur, ou à proposer un autre algorithme, puis à tester les performances de $\hat{\beta}_{M,\lambda}$ sur des données simulées ou réelles.

Bibliographie

[1] Akaike, H. (1973), Information Theory and an Extension of the Maximum Likelihood Principle, *2nd Int. Symp. on Information Theory*, Budapest: Akademia Kiado, B. N. Petrov et F. Csaki édts., pp 267-281.

[2] Alquier, P. (2008), LASSO, Iterative Feature Selection and the Correlation Selector: Oracle Inequalities and Numerical Performances, *Electron. J. Stat.*, 2, pp 1129-1152.

[3] Alquier, P. et Hebiri, M. (2008), Generalization of ℓ_1 constraint for high-dimensional regression problems, *Prépublication LPMA n. 1253*, identifiant arXiv:0811.0072, soumis.

[4] Bickel, P., Ritov, Y. et Tsybakov, A. (2007), Simultaneous Analysis of LASSO and Dantzig Selector, *Prépublication*, identifiant arXiv:0801.1095, soumis.

[5] Bunea, F. (2008), *Consistent selection via the Lasso for high dimensional approximating regression models*, IMS Collections, B. Clarke et S. Ghosal édts., pp 122-138.

[6] Bunea, F., Tsybakov, A. et Wegkamp, M. (2007), Aggregation for gaussian regression, *Ann. Statist.*, 35(4), pp 1674-1697.

[7] Candès, E. et Tao, T. (2007), The Dantzig selector: statistical estimation when p is much larger than n, *Ann. Statist.*, 35(6), pp 2313-2351.

[8] Catoni (2008), O., *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, Lecture Notes-Monograph Series vol. 56, IMS.

[9] Chapelle, O., Schölkopf, B. et Zien, A. (2006), *Semi-supervised learning*, MIT Press, Cambridge, MA.

[10] Efron, B., Hastie, T., Johnstone, I. et Tibshirani, R. (2004), Least angle regression, *Ann. Statist.*, 32(2), pp 407-199.

[11] Friedman, J., Hastie, T., Höfling, H. et Tibshirani, R. (2007), Pathwise coordinate optimization, *Ann. Appl. Statist.*, 1(2), pp 302-332.

[12] Kim, S. J., Koh, K., Lustig, M., Boyd, S. et Gorinevsky, D. (2007), An Interior-Point Method for Large-Scale ℓ_1-Regularized Least Squares, *IEEE Journ. of Sel. Topics in Signal Processing*, 1(4), pp 606-617.

[13] Schwarz, G. (1978), Estimating the Dimension of a Model, *Ann. Statist.*, 6(2), pp 461-464.

[14] Tibshirani, R. (1996), Regression shrinkage and selection via the LASSO, *JRSS-B*, 56(1), pp 267-288.

[15] Vapnik, V. (1998), *The Nature of Statistical Learning Theory*, Springer-Verlag.